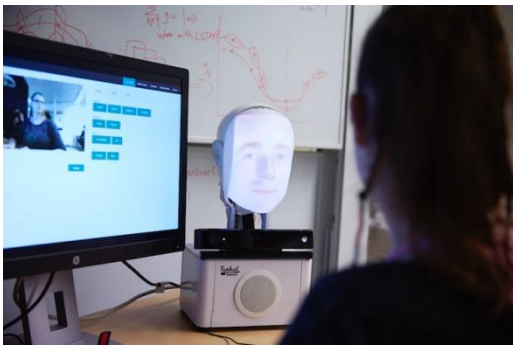
**Project title**

Explainable deep learning methods for human-human and human-robot interaction

**Project description**



Human-human interaction (HHI) relies on people's ability to mutually understand each other, often by making use of implicit signals that are physiologically embedded in human behaviour and do not require the sender's awareness, that is, honest signals. When we are engrossed in a conversation, we align with our partner: we unconsciously mimic each other, coordinate our behaviours and synchronize positive displays of emotion. This tremendously important skill, which spontaneously develops in HHI, is currently lacking in robots.

This project aims at building on advances in deep learning[1,2], and in particular on the field of Explainable Artificial Intelligence (XAI) [3,4], which offers approaches to increase the interpretability and explainability of the complex, highly nonlinear deep neural networks, to develop new machine learning-based methods that: (1) automatically analyse and predict alignment in HHI, (2) visualize and provide interpretation of regions of focus, as well as the type of used information (e.g., face expression, eye movement, body position, etc), in network's decision/prediction making to aid understanding of the alignment in HHI, (3) mine HHI data to bootstrap emotional alignment in HRI, by iteratively transferring the dynamics of behaviours learnt in HHI to HRI, thus enabling robots to align to humans, and (4) interpret and analyse the learned HHI strategies and the derived HRI alignments to both evaluate the system, and gain knowledge about subtle HHI during co-adaptive emotional alignment.

The project is a collaboration between the Uppsala Social Robotics Lab and the MIDA (Methods for Image Data Analysis) group at the Department of Information Technology, and the Uppsala Child and Baby Lab at the Department of Psychology of Uppsala University.

This interdisciplinary project will conduct ground-breaking research at the interface of social robotics, machine learning, mathematics and social psychology. New, explainable deep learning methods will be developed to visualize and interpret behavioural alignment in human-human interaction. These will enable understanding of what are the important features and actions and explain alignment in human-human interaction. This is a fundamental question of great relevance to social developmental psychology (G. Gredebäck) and social robotics (G. Castellano), as data-driven, co-adaptive methods are an emerging topic of interest for the human-robot interaction community.

The project is highly significant for the field of Explainable Artificial Intelligence (XAI) (J. Lindblad), recently identified by the European Commission's High-Level Expert Group on AI as a requirement for trustworthy AI[5]. The project results will advance the field of XAI by: (a) focusing on temporal data as well as challenges of small amount of training data – few shot learning; (b) addressing simultaneous visualisation and disentanglement/dissection of representations; (c) exploring and developing methods to analyse decision strategies; (d) exploring ways to increase human expert knowledge, as well as network knowledge, through interactive learning.

---

[1] Finn et al. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *International Conference on Machine Learning*, 2017
[2] Vaswani et al. "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017
[3] Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Int. Conf. on Computer Vision*, 2017
[4] Guidotti et al. "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)* 51(5), 93, 2019
[5] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

The project involves a number of research challenges. A prominent one is due to the project focus on spatio-temporal (video) data. Whereas recent developments has provided XAI methods applicable to static image data, the related research targeting temporal data is in its infancy and suitable solutions are yet to be developed. Creating methods for handling the multimodal data analysed in the project is particularly challenging. Inspiration may be found in the few existing initial works which are targeting temporal image data[6,7,8].

Today's techniques of XAI generally visualize *where* a network focuses its attention when making a decision, but do not explain *on what*. It is expected that motion patterns are highly important for HHI, and the project will require development of methods which can extract such types of information. However, attention is not the only target of explanation methods; we aim to reach a higher level understanding of the decision strategies learned by the models, essential for enabling human interpretation. The Spectral Relevance Analysis (SpRAy) approach[9] performs eigenvalue-based clustering to detect prediction strategies; extending this approach to temporal data will be a fruitful path of the project. Further progress towards meaningful interpretations of decision strategies will strongly benefit from their disentanglement. There exist only very initial attempts towards *disentanglement of decision strategies* and we expect that such can be developed building on the promising approaches for *representation disentanglement* based on sparse variational methods[10], which will be explored in the project. Axiomatic approaches to explanations and their evaluation[11,12,13] are particularly rewarding and will be developed as guidance for evaluation of the proposed XAI approaches, which is currently a challenge in itself.

The Uppsala Social Robotics Lab is developing technology for a storytelling scenario where a Furhat[14] robot and children collaborate to create stories. In this project, such a scenario will be extended to human-human interaction (e.g., a child collaborates with a friend or adult) to collect a multimodal data corpus using cameras, eye tracking and motion capture systems for training machine learning algorithms. The MIDA group has experience with and is actively developing methods for explainable machine learning.
The Uppsala Social Robotics Lab has already developed machine learning-based methods for data-driven human-robot interaction[15] that this project will build on. The Uppsala Child and Baby Lab has access to eye tracking and motion capture equipment to collect data and conduct experiments with children in the project. The developed methods will be tested with human-robot interaction experiments with children in schools in the city of Uppsala, which the Uppsala Social Robotics Lab already collaborates with.

The PhD candidate will be part of the Uppsala Social Robotics Lab (www.usr-lab.com) at the Division of Visual Information and Interaction of the Department of Information Technology of Uppsala University, led by Dr. Ginevra Castellano, which aims to design and develop robots that learn to interact socially with humans and bring benefits to the society we live in, for example in application areas such as education and assistive technology.
The ideal PhD candidate is a student with an MSc in Computer Science, Machine Learning, Artificial Intelligence, Robotics or related field with a broad mathematical knowledge as well as technical and programming skills. The components to be studied build on a number of mathematical techniques and the methods development involved in the project will require good command of the related areas; central are mathematical optimization and probability theory, while spectral analysis and sparse techniques have prominent roles as well. Experience and/or interest in the social sciences are also required.

---

[6] Horst et al. "Explaining the unique nature of individual gait patterns with deep learning," *Scientific reports*, 2019
[7] Hiley et al. "Explaining Temporal Information in Activity Recognition for Situational Understanding," 2019
[8] Anders et al. "Understanding patch-based learning of video data by explaining predictions," *Explainable AI*, 2019
[9] Lapuschkin et al. "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature communications* 10(1), 2019
[10] Mathieu et al. "Disentangling disentanglement in variational autoencoders," in *International Conference on Machine Learning*, 2019
[11] Lundberg, Lee. "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017
[12] Sundararajan et al. "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017
[13] Montavon. "Gradient-based vs. propagation-based explanations: an axiomatic comparison," *Explainable AI*, 2019
[14] https://www.furhatrobotics.com/
[15] Gao et al. "Fast Adaptation with Meta-Reinforcement Learning for Trust Modelling in Human-Robot Interaction," *IEEE IROS*, 2019

The PhD position is jointly funded by the Centre of Interdisciplinary Mathematics (50%) and the Division of Visual Information and Interaction (50%), Department of Information Technology, Uppsala University.

**Host institution**

Division of Visual Information and Interaction, Department of Information Technology, Uppsala University

Main advisor:

Ginevra Castellano, Uppsala Social Robotics Lab, Department of Information Technology, Uppsala University
Webpage: http://user.it.uu.se/~ginca820/

Co-advisors:

Joakim Lindblad, MIDA, Department of Information Technology, Uppsala University
Webpage: http://www.cb.uu.se/~joakim/

Gustaf Gredebäck, Uppsala Child and Baby Lab, Department of Psychology, Uppsala University
Webpage: https://psyk.uu.se/uppsala-child-and-baby-lab/people/gustaf-gredeback/